

# Weighted Difference Approximation of Value Functions for Slow-Discounting Markov Decision Processes

Yin-Lam Chow and Junjie Qin

**Abstract**—Modern applications of the theory of Markov Decision Processes (MDPs) often require frequent decision making, that is, taking an action every microsecond, second, or minute. Infinite horizon discount reward formulation is still relevant for a large portion of these applications, because actual time span of these problems can be months or years, during which discounting factors due to e.g. interest rates are of practical concern. In this paper, we show that, for such MDPs with discount rate  $\alpha$  close to 1, under a common ergodicity assumption, a weighted difference between two successive value function estimates obtained from the classical value iteration (VI) is a better approximation than the value function obtained directly from VI. Rigorous error bounds are established which in turn show that the approximation converges to the actual value function in a rate  $(\alpha\beta)^k$  with  $\beta < 1$ . This indicates a geometric convergence even if discount factor  $\alpha \rightarrow 1$ . Furthermore, we explicitly link the convergence speed to the system behaviors of the MDP using the notion of  $\epsilon$ -mixing time and extend our result to Q-functions. Numerical experiments are conducted to demonstrate the convergence properties of the proposed approximation scheme.

## I. INTRODUCTION

A large number of practical problems that involved with decision making under uncertainty can be modeled as Markov Decision Problems (MDPs). Among them, many with relatively long planning horizons are suitably casted as infinite horizon MDPs, with either discounted reward or average reward criteria [1]. While discounted reward formulation features easier-to-implement computational methods such as value iteration, in cases where the discount factor is very close to 1, it is known that the convergence for the discounted reward value iteration can be unacceptably slow. This occurs for example in communication network and computer systems applications where decisions have to be made frequently. The average reward criteria, together with their theoretical analysis and algorithmic development, were in part motivated by these observations. However, for these slow-discounting problems, the approach of first modeling the problem approximately as an average reward MDP and then solving it with corresponding algorithms (cf. Chapter 5 of [2] for more details) may give a suboptimal policy with respect to the original discounted reward criteria.

This paper provides a scheme for approximating value functions of slow-discounting MDPs. The approximation is in the form of a weighted difference between two successive value function estimates obtained from the classical VI. In particular, building from theories connecting the average reward criteria and discounted reward criteria, we demonstrate

that the approximation has a geometric convergence with an error bound of the order  $(\alpha\beta)^k$  which approaches zero even when  $\alpha \rightarrow 1$ , where  $\beta < 1$  under a common ergodicity assumption and  $k$  is the iteration count for VI. The rate parameter  $\beta$  is then characterized with the well-understood notion of  $\epsilon$ -mixing time for average reward problems.

The contributions of this paper are summarized as follows:

- We show that using a weighted difference between two successive iterates, the classical VI algorithm can be made practical even if the discount factor is arbitrarily close to one.
- We characterize the convergence of such value function approximation and discuss its relation to the notation of  $\epsilon$ -mixing time. The error bounds for the value function approximation provides novel insights on the discounted Bellman operator for ergodic MDPs, and theoretical backups for learning algorithms which may need to solve slow-discounting MDPs in its iterates<sup>1</sup>.
- We extend the above weighted difference approximation scheme to Q-functions, which is more commonly used in many reinforcement learning algorithms.

## A. Related Literature

Several methods have been proposed for solving MDPs with discount factor  $\alpha$  close to 1. Among them, splitting methods and relative value iteration (RVI) are well studied. The Gauss-Seidel VI is the most noteworthy example of splitting methods [1], which has  $(\alpha\beta^{\text{GS}})^k$  convergence, where  $\alpha\beta^{\text{GS}}$  is related to the norm of corresponding splitting matrices. However the  $\beta^{\text{GS}} < 1$  term is usually difficult to evaluate in general settings. In Section VI, the performance of our approximation scheme and Gauss-Seidel VI is compared numerically. The RVI algorithm, proposed by [4] for average reward problems and generalized to discounted reward settings by [5] and [6], is shown to have a  $(\alpha\beta^{\text{RVI}})^k$  convergence in [7]. The convergence is proved in terms of the relative value function, which is the difference between the value of each state and the value of a fixed pre-selected state, and  $\beta^{\text{RVI}} < 1$  is the second largest eigenvalue of the transition probability matrix corresponding to the optimal policy. While both the RVI and our approximation scheme are analyzed under a similar ergodicity assumption, we contrast these two approaches as follows:

- RVI is constructed to be an algorithm to obtain the relative value function, which provides sufficient in-

Y.-L. Chow and J. Qin are with the Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA. Email: {ychow, jqin}@stanford.edu.

<sup>1</sup>For example, the polynomial sample complexity bounds for reinforcement learning algorithm proposed in [3] will not be meaningful if  $\alpha \rightarrow 1$  and if classical VI is used for solving the MDP in each step.

formation to compute the optimal policy. However, to get the actual value function, one has to perform one-step policy evaluation after the algorithm converges, which requires solving a large linear system when the number of state is tremendous. Our approximation scheme estimates the value function directly, which is superior to RVI in applications such as hybrid systems where the actual value functions for each subsystem are often needed for comparison.

- The  $\beta^{\text{RVI}}$  term in the convergence rate of RVI is hard to evaluate ahead of solving the problem since it corresponds to the optimal policy. Our convergence rate can be obtained directly from the problem data beforehand.
- Our approach is both conceptually and implementation-wise simpler as its major computation is merely the classical VI.

## B. Paper Organization

The rest of the paper is organized as follows. Section II introduces the problem setup and definitions used. The approximation scheme based on weighted difference is provided in Section III, followed by a proof on its error bound. A characterization for the rate parameter  $\beta$  is derived in Section IV, based on a connection to the concept of  $\epsilon$ -mixing time. The results of a numerical experiment are given in Section VI. Finally, this paper concludes with Section VII.

## II. PROBLEM SETUP

Consider an infinite horizon discounted MDP characterized by the quintuple  $(\mathcal{S}, \mathcal{A}, R, P, \alpha)$ . Here  $\mathcal{S}$  and  $\mathcal{A}$  are finite sets representing the state space and the action space. For each  $(x, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ ,  $R_a(x, y) \in [0, R_{\max}]$  and  $P_a(x, y) \in [0, 1]$  are reward and probability of transitioning from state  $x$  to state  $y$  after taking action  $a$ , respectively. The discount rate is denoted as  $\alpha \in (0, 1)$ . In standard MDPs, the agent aims to identify a stationary policy  $\mu : \mathcal{S} \rightarrow \mathcal{A}$  that maximizes the expected discounted reward

$$\mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^t R_{\mu(x_t)}(x_t, x_{t+1}) \right].$$

Starting from each state  $x \in \mathcal{S}$ , the  $N$ -step accumulated discounted reward for policy  $\mu$  is defined as

$$V_{\mu}^N(x) = \mathbb{E} \left[ \sum_{t=1}^N \alpha^t R_{\mu(x_t)}(x_t, x_{t+1}) \middle| x_0 = x \right].$$

By Monotone Convergence Theorem, the infinite horizon value function with respect to control policy  $\mu$  is given by

$$V_{\mu}(x) = \lim_{N \rightarrow \infty} V_{\mu}^N(x) = \mathbb{E} \left[ \sum_{t=1}^{\infty} \alpha^t R_{\mu(x_t)}(x_t, x_{t+1}) \middle| x_0 = x \right]$$

and the (optimal) value function is defined by  $V^*(x) \triangleq \max_{\mu} V_{\mu}(x)$ . Similarly, we can define the state-action value function for each state-action pair  $(x, a) \in \mathcal{S} \times \mathcal{A}$  and policy  $\mu$  as  $Q_{\mu}(x, a) = \sum_{y \in \mathcal{S}} P_a(x, y)(R_a(x, y) + \alpha V_{\mu}(y))$ , and the optimal  $Q$ -function as

$$Q^*(x, a) = \sum_{y \in \mathcal{S}} P_a(x, y)(R_a(x, y) + \alpha V^*(y)). \quad (1)$$

Note that  $V^*(x) = \max_{a \in \mathcal{A}} Q^*(x, a)$  is the value function that satisfies the Bellman equation:  $V^*(x) = T[V^*](x)$ , for every  $x \in \mathcal{S}$ . The Bellman operator for discounted reward function is denoted by  $T[\cdot]$ , where

$$T[V](x) \triangleq \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y)(R_a(x, y) + \alpha V(y)). \quad (2)$$

for  $\alpha \in (0, 1)$ , and  $V : \mathcal{S} \rightarrow \mathbb{R}$  is an arbitrary function. We can write expression (1) as the Bellman equation of optimal  $Q$ -function:  $Q^*(x, a) = F[Q^*](x, a)$ ,  $\forall x \in \mathcal{S}, a \in \mathcal{A}$ , where  $F[\cdot]$  is the  $Q$ -function Bellman operator, defined as

$$F[Q](x, a) = \sum_{y \in \mathcal{S}} P_a(x, y)(R_a(x, y) + \alpha \max_{b \in \mathcal{A}} Q(y, b)),$$

for  $\alpha \in (0, 1)$ , and  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is an arbitrary function. Furthermore, let  $\mu_Q$  be a policy which satisfies  $\mu_Q \in \arg \max_{a \in \mathcal{A}} Q(x, a)$ .

Ergodicity assumptions are widely used in the analysis of stochastic optimal control and reinforcement learning [3], [8]. Motivated by identical assumptions made in the analysis of the relative value iteration algorithm for average reward MDPs (cf. Proposition 5.3.2 in [2]), we give a more quantitative characterization of the ergodicity assumption.

**Assumption II.1.** *For any admissible policy  $\pi = \{\mu_0, \mu_1, \dots, \mu_{D_{\rho}-1}\}$  and initial state  $x \in \mathcal{S}$ , there exist  $\rho \in (0, 1)$ ,  $D_{\rho} > 0$  and  $y_0 \in \mathcal{S}$  such that*

$$P_{\pi}(x_0 = x, x_{D_{\rho}} = y_0) \triangleq [P_{a_0} P_{a_1} \dots P_{a_{D_{\rho}-1}}]_{xy_0} \geq \rho, \quad (3)$$

where  $a_k = \mu_k(x_k)$ ,  $k = 0, \dots, D_{\rho} - 1$ .

## III. WEIGHTED DIFFERENCE APPROXIMATION AND ITS CONVERGENCE PROPERTIES

It is well known that there are some intrinsic relationships between maximum average reward and maximum discounted reward MDPs. As discussed in [9], for any admissible control policies, an average reward can be viewed as an orthogonal projection of the discounted reward where the relative value function is a  $(1 - \alpha)$  multiple of the residual vector. Furthermore, from Theorem 1 in [10], when  $\alpha \rightarrow 1$ , the discounted reward can be approximated by maximum average reward. However, this approximation is valid only when  $\alpha \rightarrow 1$ . Also this approach has a major drawback, as finding the optimal control policies (Blackwell optimal control policies) for discounted reward MDPs is usually computationally expensive (cf. Chapter 10 of [1] for more details). Motivated by these observations, and under Assumption II.1, this section develops a new value function approximation for discounted reward MDPs using weighted difference methods, which also arises in average reward value iteration. We also show that the error bound of this algorithm is geometric and is always smaller than the classical value iteration.

For any specific  $z \in \mathcal{S}$ , define the “gain”  $\lambda^*$  and the “bias”  $h^*$  for discounted reward MDPs:

$$h^*(x) = V^*(x) - V^*(z), \quad \lambda^* = (1 - \alpha)V^*(z).$$

By Fixed Point theorem:  $T[V^*](x) = V^*(x)$ , we have the following identity:

$$\lambda^* + h^*(x) = T[h^*](x).$$

This is analogous to the Fixed Point theorem for average reward uni-chain MDPs. Now, we define

$$\beta = (1 - \rho)^{1/D_\rho} \in (0, 1). \quad (4)$$

This term can be viewed as an improved discounted factor, and it is well defined, based on the ergodicity assumption (Assumption II.1). More discussions about  $\beta$  will be given in the next section.

Now, define the weighted difference value function approximation scheme:

**WDVF Approximation Scheme** — Given an initial value function estimate  $V_0 : \mathcal{S} \rightarrow \mathbb{R}$ , and a discounted factor  $\alpha \in (0, 1)$ , for  $k \in \{1, 2, \dots\}$ , estimate the  $(k+1)^{\text{th}}$ -step value function as follows:

$$V_{k+1}(x) = \frac{T^{k+1}[V_0](x) - \alpha T^k[V_0](x)}{1 - \alpha}, \forall x \in \mathcal{S}. \quad (5)$$

Different from the classical value iteration (which estimates the value function as  $T^k[V_0]$  at the  $(k+1)^{\text{th}}$  step), the WDVF approximation uses a normalized one-step difference:  $(T^{k+1}[V_0](x) - \alpha T^k[V_0](x)) / (1 - \alpha)$  in each updates. If we represent the  $k^{\text{th}}$ -step value function estimate in classical value iteration by

$$\bar{V}_k(x) = T^k[V_0](x),$$

the  $(k+1)^{\text{th}}$ -step WDVF approximation is equivalent to

$$V_{k+1}(x) = \frac{\bar{V}_{k+1}(x) - \alpha \bar{V}_k(x)}{1 - \alpha}, \forall x \in \mathcal{S}, \alpha \in (0, 1).$$

It is obvious that for any  $\alpha \in (0, 1)$ , if  $\bar{V}_k(x) \rightarrow \bar{V}_\infty(x) = V^*(x)$ , then  $V_{k+1}(x) \rightarrow V^*(x)$ , for any  $x \in \mathcal{S}$ . In the next theorem, we will show that the error bound of WDVF approximation converges faster than the error bound of the classical value iteration. Before getting into the details, define the following constant:

$$C_D = \max_{\ell \in \{0, 1, \dots, D_\rho - 1\}} \frac{\|T^\ell[V_0] - T^\ell[h^*]\|_d}{(\alpha\beta)^\ell} > 0 \quad (6)$$

where  $\|V\|_d = \max_{x \in \mathcal{S}} V(x) - \min_{x \in \mathcal{S}} V(x)$ .<sup>2</sup> This constant will characterize the leading coefficient of the error bound in WDVF algorithm for discounted reward problems, whose explicit formulation is provided in the following theorem.

**Theorem III.1.** For  $k \in \mathbb{Z}^+$  and any  $x \in \mathcal{S}$ , let  $V_{k+1}(x)$  be the  $(k+1)^{\text{th}}$ -step WDVF approximation obtained from equation (5). This value function approximation has the following error bound in  $\|\cdot\|_d$  semi-norm:

$$\|V_{k+1} - V^*\|_d \leq \frac{\alpha(1 + \beta)(\alpha\beta)^k}{1 - \alpha} C_D \quad (7)$$

and the following error bound for any  $x \in \mathcal{S}$ :

$$-2C_D \frac{(\alpha\beta)^k}{1 - \alpha} \leq V_{k+1}(x) - V^*(x) \leq 2C_D \frac{(\alpha\beta)^k}{1 - \alpha}. \quad (8)$$

Furthermore, let  $c(x) = T^{k-1}[V_0] - T^k[V_0]$ . Then,

$$\frac{\alpha(c(x) - \|c\|_\infty)}{1 - \alpha} \leq V_{k+1}(x) - V_k(x) \leq \frac{\alpha(c(x) + \|c\|_\infty)}{1 - \alpha}. \quad (9)$$

*Proof.* See appendix.  $\square$

<sup>2</sup>The  $\|\cdot\|_d$  notation is identical to the span-semi norm notation in equation (6.6.3) in [1].

**Remark III.2.** The difference between any two successive value function estimates in the WDVF approximation scheme is bounded. However, the sequence of value function is not monotonically increasing/decreasing.

**Remark III.3.** Similar to the relative value iteration algorithm in Section 6.6.4 in [1] and in [5] (which is namely the modified dynamic programming algorithm), the WDVF approximation is based on the normalized differences between value functions. Thus, these two methods share similar semi-norm convergence rates (the definition of  $\gamma$  in Theorem 6.6.6 in [1] is identical to  $\beta$ , when  $D_\rho = 1$ ). Nevertheless, our proposed algorithm also has a convergence rate of  $(\alpha\beta)^k$  in sup-norm, while up to the authors' knowledge, no such analysis exists for the relative value iteration algorithm.

#### IV. THE CONNECTION WITH $\epsilon$ -MIXING TIME

In the previous section, we characterize the error bound of the WDVF approximation scheme in terms of  $C_D$ ,  $\alpha$  and  $\beta$ , where  $C_D$  depends on  $\alpha$ ,  $\beta$  and the value function. The intuition behind the discounted factor  $\alpha$  is very clear. However, based on equation (4), we only know that  $\beta$  is related to the ergodicity of a Markov decision process (cf. Section 3 of [11] for details). Its explicit meaning is not well understood. In order to understand the meaning behind  $\beta$ , it is natural to study the notion of “ $\epsilon$ -mixing time” in average reward MDPs. Although we will formally define this notion later,  $\epsilon$ -mixing time can be viewed as a metric that measures the “ergodic strength” (the convergence speed of sample average reward function to relative reward function) of average reward MDPs. Intuitively  $\epsilon$ -mixing time and  $\beta$  describe similar features in a Markov decision process.

In this section, we will formulate a relationship between  $\beta$  and the  $\epsilon$ -mixing time. This in turn establishes a connection between the error bound of the WDVF approximation scheme and  $\epsilon$ -mixing time.

First, define the Bellman operator for an un-discounted reward function, similar to the case of average reward MDP:

$$\bar{T}[h](x) = \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y)(h(x) + R_a(x, y)), \quad \forall x \in \mathcal{S}.$$

Also, define  $\Pi$  to be the set of sequence of general admissible policies. The average reward MDP is given by  $\max_{\pi \in \Pi} J_\pi(x_0)$ , where

$$J_\pi(x_0) = \lim_{N \rightarrow \infty} \sup \frac{1}{N} \mathbb{E} \left[ \sum_{t=1}^N R_{\mu_t(x_t)}(x_t, x_{t+1}) \right]. \quad (10)$$

and  $\pi = \{\mu_0, \mu_1, \dots\}$ . From Proposition 5.1.1 and 5.1.2 in [2], the “lim sup” can be replaced by “lim” if we restrict  $\Pi$  to be the set of stationary admissible policies, i.e.,  $\pi = \{\mu, \mu, \dots\}$ .

From Section 5.1.3, Proposition 5.1.8 in [2], for average reward MDP, suppose the relative reward  $\lambda^* : \mathcal{S} \rightarrow \mathbb{R}$  and the bias reward  $h^* : \mathcal{S} \rightarrow \mathbb{R}$  satisfy the following pair of optimality equations:

$$\lambda^*(x) = \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) \lambda^*(y), \quad (11a)$$

$$\lambda^*(x) + h^*(x) = \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) (R_a(x, y) + h^*(y)) \quad (11b)$$



where  $\bar{\mathcal{A}}$  is the set of control actions that maximizes the first optimization problem. Then,  $\mu^*$ , which attains the maximum of these two expressions simultaneously, is the stationary optimal control policy of the average reward MDP. Furthermore, the following expression holds for any  $N' \in \mathbb{N}$ .

$$\frac{1}{N'} \mathbb{E} \left[ \sum_{k=0}^{N'-1} R_{\mu^*(x_k)}(x_k, x_{k+1}) + h^*(x_{N'}) \mid x_0 = x, \mu^* \right] - \lambda^*(x) = h^*(x)/N', \quad \forall x \in \mathcal{S}. \quad (12)$$

Thus, with  $h^*(x)$  being a finite real valued bias function obtained from expression (11), by letting  $N' \rightarrow \infty$ , we can show that  $\lambda^*(x)$  is the optimal average reward:

$$\lambda^*(x) = \lim_{N' \rightarrow \infty} \frac{1}{N'} \mathbb{E} \left[ \sum_{k=0}^{N'-1} R_{\mu^*(x_k)}(x_k, x_{k+1}) \mid x_0 = x, \mu^* \right].$$

Consider a stationary policy  $\mu$  where the Markov chain induced by  $\mu$  only has one recurrent class. We call such stationary policy a uni-chain policy. By proposition 5.2.5 in [2], if all admissible stationary policies are uni-chain, Assumption II.1 holds with  $\mu_k = \mu$ , for any  $k \in \mathbb{N}$ . Proposition 5.2.3 in [2] implies that the gain  $\lambda^*(x)$  is the same for all states. Then, the first equation in expression (11) holds trivially and  $\bar{\mathcal{A}} = \mathcal{A}$ . Thus the stationary optimal policy  $\mu^*$  can be found by the following expression:

$$\mu^*(x) \in \arg \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) (R_a(x, y) + h^*(x))$$

and  $\lambda^*$  is the optimal average reward that satisfies the fixed point theorem for average reward MDP:

$$\lambda^* + h^*(x) = \bar{T}[h^*](x), \quad \forall x \in \mathcal{S}.$$

Next, the notion of  $\epsilon$ -mixing time in a MDP is discussed. The standard notion of mixing time of a stationary control policy  $\mu$  quantifies the smallest number  $N$  of steps required to ensure that the distribution on states after  $N$  steps is within  $\epsilon$  of the stationary distribution induced by  $\mu$ . The distance between these distributions is measured by the Kullback-Leibler divergence, the variation distance, or some other standard metrics. There are well-known methods for bounding this mixing time in terms of the second eigenvalue of the transition probability matrix  $P$ , using underlying structural properties such as “conductance”. Similar to Definition 5 in [3], it turns out that we can state our results for a weaker notion of mixing time that only requires the expected discounted reward after  $N$  steps, induced by the stationary optimal control policy to approach an asymptotic reward.

**Definition IV.1.** The  $\epsilon$ -mixing time of any stationary optimal control policy,  $\mu^* \in \arg \max_{\mu} V_{\mu}(x)$ , is the smallest constant  $\tau_{\epsilon}^*$  such that for all  $N' \geq \tau_{\epsilon}^*$  and all  $x \in \mathcal{S}$ ,

$$\left| \frac{1}{N'} \mathbb{E} \left[ \sum_{k=0}^{N'-1} R_{\mu^*(x_k)}(x_k, x_{k+1}) \mid x_0 = x, \mu^* \right] - \lambda^* \right| \leq \epsilon. \quad (13)$$

Before getting to the main result of this section, we define

$$C_A = \max_{\ell \in \{0, 1, \dots, D_{\rho}-1\}} \frac{\|\bar{T}^{\ell}[V_0] - \bar{T}^{\ell}[h^*]\|_d}{(1-\rho)^{\ell/D_{\rho}}} > 0. \quad (14)$$

Similar to the definition of  $C_D > 0$ , this coefficient will characterize the constant term of an upper bound for average reward problems. The next theorem provides this upper bound in terms of the time horizon  $N'$ ,  $C_A > 0$  and  $\beta > 0$ .

Also, it gives an expression between  $\epsilon$ -mixing time and constant  $\beta \in (0, 1)$ .

**Theorem IV.2.** Let  $V_0(x) = 0$  for any  $x \in \mathcal{S}$ . Then, for any  $x \in \mathcal{S}$ , and for any  $N' \geq 1$ , there exists a constant  $C_A > 0$  such that

$$\left| \frac{1}{N'} \mathbb{E} \left[ \sum_{k=0}^{N'-1} R_{\mu^*(x_k)}(x_k, x_{k+1}) \mid x_0 = x \right] - \lambda^* \right| \leq \frac{2C_A}{N'} \frac{\beta}{1-\beta}, \quad \forall x \in \mathcal{S}. \quad (15)$$

where

$$\mu^*(x) \in \arg \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) (h(x) + R_a(x, y)).$$

Furthermore, this implies

$$\beta \geq \epsilon \tau_{\epsilon}^* / (2C_A + \epsilon \tau_{\epsilon}^*),$$

where  $\tau_{\epsilon}^*$  is the  $\epsilon$ -mixing time in Definition IV.1.<sup>3</sup>

*Proof.* For any specific  $z \in \mathcal{S}$  and  $k \in \{1, 2, \dots\}$ , define:

$$h_k(x) = \bar{T}^k[V_0](x) - \bar{T}^k[V_0](z),$$

$$\lambda_k(x) = \bar{T}^k[V_0](x) - \bar{T}^{k-1}[V_0](z), \quad \forall x \in \mathcal{S}.$$

This implies that

$$\lambda_k(x) + h_{k-1}(x) = \bar{T}[h_{k-1}](x).$$

Recall  $\|V\|_d = \max_{x \in \mathcal{S}} V(x) - \min_{x \in \mathcal{S}} V(x)$ . Similar to the arguments in Lemma III.1 for discounted reward problems, we can show that

$$\|\bar{T}^{D_{\rho}}[V^{(1)}] - \bar{T}^{D_{\rho}}[V^{(2)}]\|_d \leq (1-\rho)\|V^{(1)} - V^{(2)}\|_d.$$

We can show by induction, and fixed point theorem of average reward MDPs that

$$k\lambda^* + h^*(x) = \bar{T}^k[h^*](x).$$

Moreover, let  $k = qD_{\rho} + \ell$ , for  $\ell = \{0, 1, \dots, D_{\rho}-1\}$ , where  $q$  is the greatest common divisor of  $k$  and  $D_{\rho}$ . As in Lemma III.1, here we can also show that

$$\|\bar{T}^k[V_0] - \bar{T}^k[h^*]\|_d \leq (\beta)^{qN} \|\bar{T}^{\ell}[V_0] - \bar{T}^{\ell}[h^*]\|_d \leq C_A \beta^k.$$

Following similar derivations as in Lemma III.1, the above results further imply that  $\|h_k - h^*\|_{\infty} \leq C_A \beta^k$  and

$$\begin{aligned} \|\bar{T}^k[V_0] - \bar{T}^{k-1}[V_0] - \lambda^*\|_{\infty} &= \|\lambda_k - \lambda^*\|_{\infty} \\ &\leq 2\|h_k - h^*\|_{\infty} \leq 2C_A \beta^k. \end{aligned}$$

Furthermore, by a telescoping sum,

$$\begin{aligned} \left| \frac{\bar{T}^N[V_0](x) - V_0(x)}{N} - \lambda^* \right| &\leq \sum_{k=1}^N \frac{\|\bar{T}^k[V_0] - \bar{T}^{k-1}[V_0] - \lambda^*\|_{\infty}}{N} \\ &\leq \frac{2C_A}{N} \sum_{k=1}^N \beta^k = \frac{2C_A}{N'} \frac{\beta(1-\beta^{N'})}{1-\beta} \leq \frac{2C_A}{N'} \frac{\beta}{1-\beta} \end{aligned}$$

for any  $x \in \mathcal{S}$ . Since  $V_0(x) = 0$  for all  $x \in \mathcal{S}$ , the above result implies expression (15). Now, for  $N_0 = 2C_A\beta/(\epsilon(1-\beta))$ , one obtains

$$\left| \frac{\mathbb{E} \left[ \sum_{k=0}^{N'-1} R_{\mu^*(x_k)}(x_k, x_{k+1}) \mid x_0 = x \right]}{N'} - \lambda^* \right| \leq \epsilon$$

for any  $N' \geq N_0$ . Then, based on the definition of  $\epsilon$ -mixing time in Definition IV.1, we conclude that  $2C_A\beta/(\epsilon(1-\beta)) \geq \tau_{\epsilon}^*$  and  $\beta \geq \epsilon \tau_{\epsilon}^* / (2C_A + \epsilon \tau_{\epsilon}^*)$ .  $\square$

<sup>3</sup>Proof of this result is omitted in this conference version and can be found at [web.stanford.edu/~ychow](http://web.stanford.edu/~ychow).

Now, we are in position to give a relationship between the number of steps needed for convergence of  $\mathcal{WDVF}$  approximation and  $\epsilon$ -mixing time  $\tau_\epsilon^*$ . Given a constant  $\theta > 0$ . From Lemma III.1, the condition  $\|V_k - V^*\|_\infty \leq \theta$  holds if

$$\frac{2C_D(\alpha\beta)^{k-1}}{1-\alpha} \leq \theta \iff k \geq \log\left(\frac{\theta(1-\alpha)}{2C_D}\right) / \log(\alpha\beta) + 1.$$

From the  $\beta$  bound given by Theorem IV.2, we know that, if the number of steps is given by the following expression:

$$k \geq C_\theta \triangleq \max\left\{\frac{\log(\theta(1-\alpha)/(2C_D))}{\log(\alpha\epsilon\tau_\epsilon^*/(2C_A + \epsilon\tau_\epsilon^*))} + 1, 1\right\}, \quad (16)$$

where  $\tau_\epsilon^*$  is the  $\epsilon$ -mixing time and  $C_D, C_A$  are given by equations (6) and (14) respectively, then  $\|V_k - V^*\|_\infty \leq \theta$  is guaranteed. We summarize this result as follows:

**Theorem IV.3.** *Let  $V_k(x)$  be the  $k^{\text{th}}$   $\mathcal{WDVF}$  approximation obtained from equation (5), for any  $x \in \mathcal{S}$ . The number of steps required for  $\|V_k - V^*\|_\infty \leq \theta$  is at least  $C_\theta$ .*

## V. MODIFIED Q-VALUE ITERATION

In this section, we study the convergence properties of modified  $Q$ -value iteration. First, define the following algorithm for modified  $Q$ -value iteration:

**$\mathcal{WDQVF}$  Approximation Scheme** — Given an initial value function estimate  $V_0 : \mathcal{S} \rightarrow \mathbb{R}$ , and a discounted factor  $\alpha \in (0, 1)$ . Let  $Q_0(x, a)$  be the following initial  $Q$ -function estimate:

$$Q_0(x, a) = V_0(x), \quad \forall (x, a) \in \mathcal{S} \times \mathcal{A}.$$

For  $k \in \{1, 2, \dots\}$ , update the  $(k+1)^{\text{th}}$ -step  $Q$ -function estimate as follows:

$$Q_{k+1}(x, a) = \frac{F^{k+1}[Q_0](x, a) - \alpha F^k[Q_0](x, a)}{1 - \alpha} \quad (17)$$

for any  $(x, a) \in \mathcal{S} \times \mathcal{A}$ .

By using the error bound result for modified value iteration from the  $\mathcal{WDVF}$  approximation, we can prove a similar error bound for  $\mathcal{WDQVF}$  approximation. This result is summarized in the following theorem.

**Theorem V.1.** *Let  $\{Q_k\}$  be a sequence of  $Q$ -value function estimates generated by the  $\mathcal{WDQVF}$  approximation scheme. Then, the following expression holds for any  $(x, a) \in \mathcal{S} \times \mathcal{A}$ :*

$$|Q_k(x, a) - Q^*(x, a)| \leq 2\alpha C_D \frac{(\alpha\beta)^{k-2}}{1-\alpha} = O((\alpha\beta)^k)$$

*Proof.* Based on the definitions of  $T[\cdot]$  and  $F[\cdot]$ , we know that  $\max_{a \in \mathcal{A}} F[Q_0](x, a) = T[V_0](x)$ . By repeating the above analysis, we can show by induction that  $\max_{a \in \mathcal{A}} F^k[Q_0](x, a) = T^k[V_0](x)$ ,  $\forall k \in \mathbb{N}$ . We will use the error bound result in the  $\mathcal{WDVF}$  approximation scheme to show a similar error bound for the  $\mathcal{WDQVF}$  approximation scheme. First, let

$$\bar{Q}(x, a) = T^k[V_0](x), \quad \forall (x, a) \in \mathcal{S} \times \mathcal{A}.$$

By applying  $F[\cdot]$  to the above equation, it implies for any  $(x, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$\begin{aligned} F[T^k[V_0]](x, a) &= F[\bar{Q}](x, a) \\ &= \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha \max_{b \in \mathcal{A}} \bar{Q}(y, b) \right) \\ &= \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha T^k[V_0](y) \right) \\ &= \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha \max_{b \in \mathcal{A}} F^k[Q_0](y, b) \right) = F^{k+1}[Q_0](x, a). \end{aligned}$$

Now, expression (21) and (22) imply

$$\begin{aligned} -\frac{2\|h_k - h^*\|_\infty}{1-\alpha} + T^k[V_0](x) &\leq V^*(x) \\ -\frac{T^k[V_0](x) - T^{k+1}[V_0](x)}{1-\alpha} &\leq \frac{2\|h_k - h^*\|_\infty}{1-\alpha} + T^k[V_0](x). \end{aligned}$$

By applying  $F[\cdot]$  to the above inequality, and noting that

$$F[Q + c](x, a) = F[Q](x, a) + \alpha c,$$

we know that for any  $(x, a) \in \mathcal{S} \times \mathcal{A}$ ,

$$\begin{aligned} &-\frac{2\alpha\|h_k - h^*\|_\infty}{1-\alpha} + F^{k+1}[Q_0](x, a) \\ &\leq F\left[V^*(x) + \frac{T^k[V_0](x) - T^{k+1}[V_0](x)}{1-\alpha}\right] \\ &\leq \frac{2\alpha\|h_k - h^*\|_\infty}{1-\alpha} + F^{k+1}[Q_0](x, a) \end{aligned} \quad (18)$$

Furthermore, by recalling  $\max_{a \in \mathcal{A}} Q^*(x, a) = V^*(x)$ , we obtain the following expressions:

$$\begin{aligned} &F\left[V^*(x) + \frac{T^k[V_0](x) - T^{k+1}[V_0](x)}{1-\alpha}\right] \\ &= \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha \max_{a \in \mathcal{A}} \left\{ V^*(y) + \frac{T^k[V_0](y) - T^{k+1}[V_0](y)}{1-\alpha} \right\} \right) \\ &= Q^*(x, a) + \alpha \sum_{y \in \mathcal{S}} P_a(x, y) \frac{T^k[V_0](y) - T^{k+1}[V_0](y)}{1-\alpha} \\ &= \frac{1}{1-\alpha} \left( \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha \max_{b \in \mathcal{A}} F^k[Q_0](y, b) \right) \right. \\ &\quad \left. - \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha \max_{b \in \mathcal{A}} F^{k+1}[Q_0](y, b) \right) \right) + Q^*(x, a) \\ &= Q^*(x, a) + \frac{1}{1-\alpha} \left( F^{k+1}[Q_0](x, a) - F^{k+2}[Q_0](x, a) \right). \end{aligned}$$

Thus, by combining all arguments, expression (18) implies

$$\begin{aligned} &-\frac{2\alpha\|h_k - h^*\|_\infty}{1-\alpha} \\ &\leq Q^*(x, a) - \left( \frac{F^{k+2}[Q_0](x, a)}{1-\alpha} - \frac{\alpha}{1-\alpha} F^{k+1}[Q_0](x, a) \right) \\ &\leq \frac{2\alpha\|h_k - h^*\|_\infty}{1-\alpha} \end{aligned}$$

Now, by putting the result:  $\|h_k - h^*\|_\infty \leq C_D(\alpha\beta)^k$  to the above expression, the error bound proof for the  $\mathcal{WDQVF}$  approximation scheme is completed.  $\square$

## VI. NUMERICAL EXPERIMENT

Consider 100 Monte Carlo samples of randomly generated 100-state-6-action MDPs with  $\mathcal{S} = \{1, 2, \dots, 100\}$ ,  $\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$ ,  $\alpha = 0.995$ . The reward functions are randomly generated with  $R_{\max} = 1$ . For simplicity each reward function is assumed to be  $y$ -independent, that is,  $R_a(x, y) = R_a(x)$  along  $y \in \mathcal{S}$ . The transition probabilities induced by each actions are randomly generated with ergodic strength of at least 0.1 ( $\rho = 0.1$  and  $D_\rho = 1$ ). This further implies the improved discount factor  $\beta$  equals to 0.9.<sup>4</sup> We want to compare the performance between the classical value iteration, Gauss-Seidel value iteration and the  $\mathcal{WDVF}$  approximation scheme. Recall that the error bound for value iteration is given by  $R_{\max}\alpha^k/(1-\alpha)$ . From Theorem III.1, the error bound for  $\mathcal{WDVF}$  approximation is given by  $2C_D(\alpha\beta)^{k-1}/(1-\alpha)$ . From Proposition 6.3.8 in [1], the error bound of Gauss-Seidel value iteration is given by  $R_{\max}(\alpha\beta^{\text{GS}})^k/(1-\alpha)$ , where  $\beta^{\text{GS}} < 1$  can be calculated using the matrix regular splitting method depicted in Theorem 6.3.4 of [1].

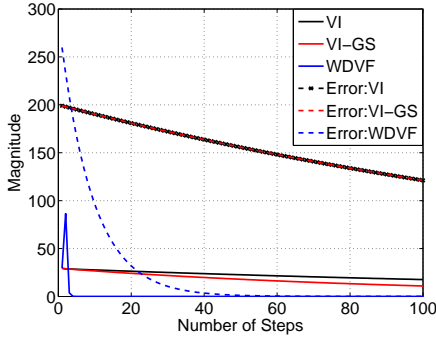


Fig. 1: Mean of  $\|V_k - V^*\|_\infty$  across Monte Carlo runs.

Figure 1 compares the error bound and speed of convergence of the  $\mathcal{WDVF}$  approximation scheme, Gauss-Seidel value iteration and the classical value iteration. The stopping criterion of this experiment is:  $\|V_k - V^*\|_\infty \leq 10^{-5}$ . On average, it is observed that  $\mathcal{WDVF}$  approximation takes 92 iterations (standard deviation: 13 iterations) to converge, while Gauss-Seidel value iteration and classical value iteration take 2936 iterations (standard deviation: 197 iterations) and 3551 iterations (standard deviation: 172 iterations) to converge respectively. As illustrated in Theorem III.1, the error bound of  $\mathcal{WDVF}$  approximation is in the order of  $(\alpha\beta)^k = 0.8996^k$ , while the error bound of the classical value iteration and Gauss-Seidel value iteration are in the order of  $0.995^k$  (as  $\alpha\beta^{\text{GS}} \simeq \alpha$  numerically in our experiment). This numerical example demonstrates that, when  $\alpha \rightarrow 1$ , both classical value iteration and Gauss Seidel may encounter slow convergence issues, while the convergence for the  $\mathcal{WDVF}$  approximation depends on  $\beta$ .

<sup>4</sup>The explicit formulations of the reward functions and transition probabilities can be found in the author's website.

## VII. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel weighted difference value function estimation scheme for discounted reward MDPs. We have shown that this approximation has an error bound of order  $(\alpha\beta)^k$ ,  $\beta \in (0, 1)$ , which decays faster than the error bound of classical value iteration (in order of  $\alpha^k$ ). We also characterize the improved convergence factor  $\beta$  and the speed of convergence of this new approximation using  $\epsilon$ -mixing time. This characterization explicitly links the convergence speed of weighted difference value function estimation to the system behaviors of the MDP. Furthermore, we also extend the above method to find optimal  $Q$ -function. The above theoretical result is verified by a numerical experiment. Notice that while Assumption II.1 can be justified via Schweitzer's transformation [12] in average reward MDPs, similar transformation does not work under the discounted reward settings. Eliminating the restrictions due to the ergodicity assumption will be left as future work.

### ACKNOWLEDGEMENT

The authors would like to thank Professor Benjamin Van Roy for invaluable discussions.

### REFERENCES

- [1] M. Puterman. *Markov Decision Process, Discrete Stochastic Dynamic Programming*. Wiley, 2005.
- [2] D. P. Bertsekas. *Dynamic Programming and Optimal Control: Approximate Dynamic Programming, Volume 2, 4th Edition*. Athena Scientific, 2012.
- [3] M. Kearns and S. Singh. Near-Optimal Reinforcement Learning in Polynomial Time. *Machine Learning*, 49(2):209–232, 2002.
- [4] D. J. White. Dynamic Programming, Markov Chains, and the Method of Successive Approximations. *Journal of Mathematical Analysis and Applications*, 6(3):373–376, 1963.
- [5] J. MacQueen. A Modified Dynamic Programming Method for Markovian Decision Problems. *Journal of Mathematical Analysis and Applications*, 14(1):38–43, 1966.
- [6] A. R. Odoni. On Finding the Maximal Gain for Markov Decision Processes. *Operations Research*, 17(5):857–860, 1969.
- [7] T. E. Morton. On the Asymptotic Convergence Rate of Cost Differences for Markovian Decision Processes. *Operations Research*, 19(1):244–248, 1971.
- [8] R. I. Brafman and M. Tennenholtz. R-MAX: A General Polynomial Time Algorithm for Near-Optimal Reinforcement Learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.
- [9] J. N. Tsitsiklis and B. Van Roy. On Average Versus Discounted Reward Temporal-Difference Learning. *Machine Learning*, 49(2-3):179–191, 2002.
- [10] S. Kakade. Optimizing Average Reward Using Discounted Rewards. In *Proc. of the 14th Annual Conference on Computational Learning Theory and 5th European Conference on Computational Learning Theory*, pages 605–615, 2001.
- [11] T. E. Morton and W. E. Wecker. Discounting, Ergodicity and Convergence for Markov Decision Processes. *Management Science*, 23(8):890–900, 1977.
- [12] R. Cavazos-Cadena. A Note on the Convergence Rate of the Value Iteration Scheme in Controlled Markov Chains. *Systems & control letters*, 33(4):221–230, 1998.

### APPENDIX

**Proof of Theorem III.1.** Let  $V^{(1)}$  and  $V^{(2)}$  be two arbitrary functions that maps  $\mathcal{S}$  to  $\mathbb{R}$ . We define  $V_k^{(1)}(x) = T^k[V^{(1)}](x)$  and  $V_k^{(2)}(x) = T^k[V^{(2)}](x)$  for any  $x \in \mathcal{S}$  and for  $k \in \{0, \dots, D_\rho\}$ . We also define two sequences of optimal policies,  $\pi^{(j)} = \{\mu_0^{(j)}, \mu_1^{(j)}, \dots\}$ , for  $j \in \{1, 2\}$ , where  $\mu_k^{(j)}(x) \in$

$\arg \max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha V_k^{(j)}(y) \right)$ .

For any sequence of state feedback control policies  $\pi = \{\mu_0, \mu_1, \dots\}$ , define the following event:  $\mathcal{H}(x, \pi) = \{x_0 = x, a_i = \mu_i(x_i), \forall i\}$ , where  $\{x_j\}_{j \in \mathbb{Z}^+}$  is a Markov chain induced by control policy  $\pi$  with  $x_0 = x$ . By substituting the sequences of optimal policies to value function  $V_{D_\rho}^{(j)}$ , one notices that for  $j \in \{1, 2\}$ , and for any  $x \in \mathcal{S}$ ,

$$V_{D_\rho}^{(j)}(x) = \mathbb{E} \left[ \sum_{i=0}^{D_\rho-1} \alpha^i R_{a_i}(x_i, x_{i+1}) + \alpha^{D_\rho} V^{(j)}(x_{D_\rho}) \mid \mathcal{H}(x, \pi^{(j)}) \right].$$

By considering the difference between  $V_{D_\rho}^{(1)}(x)$  and  $V_{D_\rho}^{(2)}(x)$ , we get

$$\begin{aligned} & V_{D_\rho}^{(1)}(x) - V_{D_\rho}^{(2)}(x) \\ &= \mathbb{E} \left[ \sum_{i=0}^{D_\rho-1} \alpha^i R_{a_i}(x_i, x_{i+1}) + \alpha^{D_\rho} V^{(1)}(x_{D_\rho}) \mid \mathcal{H}(x, \pi^{(1)}) \right] \\ &- \mathbb{E} \left[ \sum_{i=0}^{D_\rho-1} \alpha^i R_{a_i}(x_i, x_{i+1}) + \alpha^{D_\rho} V^{(2)}(x_{D_\rho}) \mid \mathcal{H}(x, \pi^{(2)}) \right] \\ &\geq \mathbb{E} \left[ \sum_{i=0}^{D_\rho-1} \alpha^i R_{a_i}(x_i, x_{i+1}) + \alpha^{D_\rho} V^{(1)}(x_{D_\rho}) \mid \mathcal{H}(x, \pi^{(2)}) \right] \\ &- \mathbb{E} \left[ \sum_{i=0}^{D_\rho-1} \alpha^i R_{a_i}(x_i, x_{i+1}) + \alpha^{D_\rho} V^{(2)}(x_{D_\rho}) \mid \mathcal{H}(x, \pi^{(2)}) \right] \\ &= \mathbb{E} \left[ \alpha^{D_\rho} (V^{(1)}(x_{D_\rho}) - V^{(2)}(x_{D_\rho})) \mid \mathcal{H}(x, \pi^{(2)}) \right]. \end{aligned}$$

The first inequality is due to the fact that for any  $k \in \mathbb{Z}^+$ ,  $\mu_k^{(2)}(x)$  is a feasible solution to the optimization problem  $\max_{a \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) \left( R_a(x, y) + \alpha V_k^{(1)}(y) \right)$ ,

where  $\mu_k^{(1)}(x)$  is an optimal solution of this problem, for every  $x \in \mathcal{S}$ . By Assumption II.1, this further implies that

$$\begin{aligned} & (V_{D_\rho}^{(1)}(x) - V_{D_\rho}^{(2)}(x)) / \alpha^{D_\rho} \\ &\geq \sum_{y \in \mathcal{S}} \mathbb{P}_{\pi^{(2)}}(x_0 = x, x_{D_\rho} = y) (V^{(1)}(y) - V^{(2)}(y)) \\ &\geq [(1 - \rho) \min_{y \in \mathcal{S}} \{V^{(1)}(y) - V^{(2)}(y)\} + \rho(V^{(1)}(y_0) - V^{(2)}(y_0))], \end{aligned}$$

where  $y_0 \in \mathcal{S}$  is the state defined in Assumption II.1.

Similarly, by a symmetric argument, we can also prove that

$$\begin{aligned} & \frac{1}{\alpha^{D_\rho}} \max_{y \in \mathcal{S}} \left\{ T^{D_\rho}[V^{(1)}](y) - T^{D_\rho}[V^{(2)}](y) \right\} \\ &\leq [(1 - \rho) \max_{y \in \mathcal{S}} \{V^{(1)}(y) - V^{(2)}(y)\} + \rho(V^{(1)}(y_0) - V^{(2)}(y_0))]. \end{aligned}$$

Thus, by these inequalities and the definitions of  $\|T^{D_\rho}[V^{(1)}] - T^{D_\rho}[V^{(2)}]\|_d$ ,  $\|V^{(1)} - V^{(2)}\|_d$ , we can show that the following  $D_\rho$ -step contraction property holds:

$$\|T^{D_\rho}[V^{(1)}] - T^{D_\rho}[V^{(2)}]\|_d \leq (\alpha\beta)^{D_\rho} \|V^{(1)} - V^{(2)}\|_d.$$

By mathematical induction and the definitions of  $\lambda^*$ ,  $h^*$ , it can be easily shown that

$$\sum_{i=0}^{k-1} \alpha^i \lambda^* + h^*(x) = T^k[h^*](x), \quad \forall x \in \mathcal{S}. \quad (19)$$

Consider the expression:  $\left\| T^k[V_0] - \sum_{i=0}^{k-1} \alpha^i \lambda^* - h^* \right\|_d$ . By writing  $k = qD_\rho + \ell$ ,  $\ell = \{0, 1, \dots, D_\rho - 1\}$ , where the nonnegative integer  $q$  is the greatest common divisor of  $k$  and  $D_\rho$ , from expression (19), we obtain the following relationship:

$$\begin{aligned} & \left\| T^k[V_0] - \sum_{i=0}^{k-1} \alpha^i \lambda^* - h^* \right\|_d = \|T^k[V_0] - T^k[h^*]\|_d \quad (20) \\ &\leq (\alpha\beta)^{qD_\rho} \|T^\ell[V_0] - T^\ell[h^*]\|_d \leq C_D(\alpha\beta)^k. \end{aligned}$$

Note that  $T^k[h^*](x) = T^k[V^*](x) - \alpha^k V^*(z) = V^*(x) - \alpha^k V^*(z)$ . From Section 6.6.1 in [1], one also obtains  $\|u + v\|_d \leq \|u\|_d + \|v\|_d$ ,  $\| -u \|_d = \|u\|_d$ ,  $\|ku\|_d = |k| \|u\|_d$  and  $\|u + k\|_d = \|u\|_d$  for any scalar  $k$ . Therefore, the above expression implies  $\|T^k[V_0] - T^k[h^*]\|_d = \|T^k[V_0] - V^*\|_d \leq C_D(\alpha\beta)^k$  and

$$\begin{aligned} \|V_{k+1} - V^*\|_d &= \left\| \frac{(T^{k+1}[V_0] - V^*) - \alpha(T^k[V_0] - V^*)}{1 - \alpha} \right\|_d \\ &\leq (\|T^{k+1}[V_0] - V^*\|_d + \alpha\|T^k[V_0] - V^*\|_d) / (1 - \alpha) \\ &\leq (\alpha\beta)^k ((\alpha\beta) + \alpha) C_D / (1 - \alpha). \end{aligned}$$

This implies that the error bound in expression (7) holds.

Next, we will show the error bound in expression (8). Define the following quantities that estimate the gain and bias in the  $k^{\text{th}}$  step:

$$h_k(x) = T^k[V_0](x) - T^k[V_0](z),$$

$$\lambda_k(x) = T^k[V_0](x) - T^{k-1}[V_0](x) + (1 - \alpha)T^{k-1}[V_0](z)$$

where  $z \in \mathcal{S}$  is an arbitrary reference state. By simple calculations, the above expressions imply  $\lambda_{k+1}(x) + h_k(x) = T[h_k](x)$ . It can be easily seen that  $h^*(z) = V^*(z) - V^*(z) = 0$  and

$$\begin{aligned} |h_k(x) - h^*(x)| &= |T^k[V_0](x) - T^k[V_0](z) - h^*(x) + h^*(z)| \\ &= |T^k[V_0](x) - \sum_{i=0}^{k-1} \alpha^i \lambda^* - h^*(x) - [T^k[V_0](z) - \sum_{i=0}^{k-1} \alpha^i \lambda^* - h^*(z)]| \\ &\leq \|T^k[V_0] - \sum_{i=0}^{k-1} \alpha^i \lambda^* - h^*\|_d \leq C_D(\alpha\beta)^k. \end{aligned}$$

Thus, the above inequality implies

$$\|h^* - h_k\|_\infty = \max_{x \in \mathcal{S}} |h^*(x) - h_k(x)| \leq C_D(\alpha\beta)^k.$$

Next, we know from the contraction property of  $T[\cdot]$  that

$$\begin{aligned} & \lambda_{k+1}(x) + h_k(x) - (\lambda^* + h^*(x)) = T[h_k](x) - T[h^*](x) \\ &\leq \max_{b \in \mathcal{A}} \alpha \sum_{y \in \mathcal{S}} P_b(x, y) |h_k(y) - h^*(y)| \leq \alpha \|h_k - h^*\|_\infty. \end{aligned}$$

By using the definitions of  $\lambda_{k+1}(x)$ ,  $h_k(x)$ ,  $\lambda^*$  and  $h^*(x)$ , the above expression implies

$$\begin{aligned} & T^{k+1}[V_0](x) - T^k[V_0](x) + (1 - \alpha)(T^k[V_0](z) - V^*(z)) \\ &\quad + h_k(x) - h^*(x) \leq \alpha \|h_k - h^*\|_\infty, \end{aligned}$$

which further implies

$$\begin{aligned} & T^{k+1}[V_0](x) - T^k[V_0](x) + (1 - \alpha)(T^k[V_0](z) - V^*(z)) \\ &\leq (1 + \alpha) \|h_k - h^*\|_\infty. \end{aligned}$$

By inserting

$$T^k[V_0](z) - V^*(z) = T^k[V_0](x) - h_k(x) - (V^*(x) - h^*(x))$$

to the above expression, we get

$$\begin{aligned} & T^{k+1}[V_0](x) - T^k[V_0](x) + (1 - \alpha)(T^k[V_0](x) - V^*(x)) \\ &\quad - (1 - \alpha)(h_k(x) - h^*(x)) \leq (1 + \alpha) \|h_k - h^*\|_\infty. \end{aligned}$$

This implies

$$\begin{aligned} & T^{k+1}[V_0](x) - T^k[V_0](x) + (1 - \alpha)(T^k[V_0](x) - V^*(x)) \\ &\leq (1 + \alpha) \|h_k - h^*\|_\infty + (1 - \alpha)(h_k(x) - h^*(x)) \\ &\leq 2 \|h_k - h^*\|_\infty. \end{aligned}$$

By combining all inequalities, we get

$$\frac{T^{k+1}[V_0](x) - \alpha T^k[V_0](x)}{1 - \alpha} - V^*(x) \leq \frac{2 \|h_k - h^*\|_\infty}{1 - \alpha}. \quad (21)$$

Similarly, by noting that

$$\begin{aligned} & T[h_k](x) - T[h^*](x) \\ & \geq -\alpha \max_{b \in \mathcal{A}} \sum_{y \in \mathcal{S}} P_a(x, y) |h_k(y) - h^*(y)| \geq -\alpha \|h_k - h^*\|_\infty \end{aligned}$$

and applying analogous arguments as in the derivation of inequality (21), we get

$$\frac{T^{k+1}[V_0](x) - \alpha T^k[V_0](x)}{1 - \alpha} - V^*(x) \geq -2 \frac{\|h_k - h^*\|_\infty}{1 - \alpha}. \quad (22)$$

Now, since  $\|h_k - h^*\|_\infty \leq C_D(\alpha\beta)^k$ , the definition of  $V_{k+1}(x)$ , expression (21) and (22) imply expression (8) holds for all  $x \in \mathcal{S}$ . This completes the first part of the proof.

Finally, we will show expression (9) holds. For any  $k \in \mathbb{Z}^+$ , the  $\mathcal{WDVF}$  approximation can be re-written as

$$V_{k+1}(x) = \frac{T^{k+1}[V_0](x) - T^k[V_0](x)}{1 - \alpha} + T^k[V_0](x).$$

Thus, we know that

$$\begin{aligned} V_{k+1}(x) - V_k(x) &= \frac{T^{k+1}[V_0](x) + T^{k-1}[V_0](x) - 2T^k[V_0](x)}{1 - \alpha} \\ &\quad + T^k[V_0](x) - T^{k-1}[V_0](x) \\ &\leq \frac{\alpha(\|T^k[V_0] - T^{k-1}[V_0]\|_\infty + T^{k-1}[V_0](x) - T^k[V_0](x))}{1 - \alpha}. \end{aligned}$$

The first inequality is implied by the fact that  $T[\cdot]$  is a  $\alpha$ -contraction mapping:

$$T^{k+1}[V_0](x) - T^k[V_0](x) \leq \alpha \|T^k[V_0] - T^{k-1}[V_0]\|_\infty, \quad \forall x \in \mathcal{S}.$$

On the other hand, we can also show that

$$\begin{aligned} & V_{k+1}(x) - V_k(x) \\ & \geq \frac{\alpha(-\|T^k[V_0] - T^{k-1}[V_0]\|_\infty + T^{k-1}[V_0](x) - T^k[V_0](x))}{1 - \alpha}. \end{aligned}$$

by analogous arguments. This completes the second part of the proof.